DOCUMENT RESUME

ED 480 042 CG 032 615

AUTHOR	Ellis, Barbara B.; Raju, Nambury S.
TITLE	Test and Item Bias: What They Are, What They Aren't, and How To Detect Them.
PUB DATE	2003-08-00
NOTE	12p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Educational Assessment; *Educational Testing; Evaluation Methods; *Item Bias; Psychometrics; Student Evaluation; *Test Bias; *Test Construction; *Testing Problems

ABSTRACT

This chapter briefly describes some of the methods that test developers and psychometricians have devised to identify item and test bias and some of the challenges they still face. Although it may not be reasonable for classroom teachers to use these methods on a day-to-day basis in constructing tests, the authors argue that it is important for readers to know that these methods are widely used by researchers, professional test developers, and state agencies that develop standardized tests of student achievement. (Contains 12 references.) (GCP)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 480 042

Test and Item Bias: What They Are, What They Aren't, and How to Detect Them

By
Barbara B. Ellis
Nambury S. Raju

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

CG032615

Chapter 7
Test and Item Bias
What They Are, What They Aren't, and How to Detect Them

Barbara B. Ellis & Nambury S. Raju

When laypersons refer to a test as biased, they usually think of the test as measuring different test takers in different ways. For example, when someone says that a test of cognitive ability is biased against a group of test takers, the assumption is that the test systematically assesses something other than cognitive ability. Laypersons commonly assume that because there are consistent differences in obtained cognitive ability for Asians versus Whites, and for Whites versus Blacks, on tests of cognitive ability, the tests must be biased. The implication is that these tests are more difficult for some test takers because the test is composed of items written in a manner that does not account for cultural differences between these groups of test takers.

In contrast, no one would argue that a yardstick is a biased measure of the construct we refer to as height. We do not question that a yardstick measures height for everyone in the same manner. As a measurement instrument, we do not suspect that a yardstick is influenced by factors other than the construct it is intended to measure— height. Thus, when a yardstick is used to measure two individuals who are equal in height but who differ in gender or ethnicity, they can be expected to have the same "score" in terms of inches of height. We feel comfortable saying the yardstick is an unbiased measurement instrument. Just because the yardstick is unbiased, however, does not mean that, on average, one group will be the same in height as another. On average, women are likely to be somewhat shorter than men, and Hispanics and Asians are likely to be somewhat shorter than Caucasian Americans or African Americans. In other words, an unbiased measurement instrument does not necessarily imply that different groups will have the same average scores on the construct assessed—groups do differ in average cognitive ability just as they differ in height. (Chapter 10 addresses socioeconomic and cultural factors that may interfere with test performance.)

Likewise, when we assess a psychological construct (e.g., cognitive ability), we would like to obtain test scores that are not

influenced by factors that are irrelevant to the construct that the test intends to measure (AERA, APA, & NCME, 1999). For example, scores should not be influenced by factors such as the test takers' group membership but should measure individuals from different groups in the same manner.

Imagine a test designed to measure the construct of mechanical reasoning. If test takers are equal in mechanical reasoning ability, even if they come from different groups, we would expect them to have the same probability of answering an item correctly. A question with these characteristics would be considered unbiased. If, on the other hand, two test takers are equal in mechanical reasoning ability but do not have the same chance of answering a particular mechanical reasoning item correctly, we would question whether this item is measuring mechanical reasoning in the same manner for both examinees. The test takers in this example are, by definition, equal in mechanical reasoning and should have the same probability of a correct response. In that case, we may conclude that the test question is not measuring mechanical ability in the same fashion for these two test takers, that is, the item is biased. If the test were composed of many items like those just described, and if these items always functioned such that one test taker had a higher probability, and the other a lower probability, of answering correctly, we would consider the test to be biased as well. On the other hand, if our test were composed of items like those first described (i.e., test takers who are equal in mechanical reasoning, regardless of group membership, have the same probability of answering correctly), we would consider the test unbiased. Like the yardstick, the latter test is functioning in the same fashion for all test takers; however, this does not preclude there being differences in average mechanical reasoning scores at the group level.

For test developers and psychometricians, the problem becomes one of developing methods that can be used to support the assumption that test takers are equal in the construct being assessed. Once that is accomplished, we can look at the likelihood that examinees who are equivalent in the psychological construct assessed, but who come from different groups, have the same probability of answering a test item correctly (i.e., have the same expected score). If that is the case, we can conclude that the item is measuring in an equivalent fashion for both groups, that is, the item is unbiased. At the test level, we may conclude that a test is unbiased in two ways. Obviously, if a test does not contain any biased items, we would conclude that the test is unbiased. In addition, if we find some items that function against a particular group,

but other items function in favor of that group such that the effects of the biased items cancel each other out, the test may be unbiased at the test level (not the item level).

In the remainder of this chapter we briefly describe some of the methods that test developers and psychometricians have devised to identify item and test bias and some of the challenges they still face. Although it may not be reasonable for classroom teachers to use these methods on a day-to-day basis in constructing tests, it is important for readers to know that these methods are widely used by researchers, professional test developers, and state agencies that develop standardized tests of student achievement. Finally, we would like readers to know and understand that if groups differ in test scores, this does not necessarily mean that a test is biased. If we can determine that a test is composed of unbiased items (or that biased items balance out at the test level), we may conclude that the test is unbiased. As in our yardstick example, groups may differ in their test scores, even if the test is unbiased. It is, however, necessary to identify item and test bias prior to comparing group test scores. Without this assessment, we cannot be sure if scores at the group level differ due to item bias or real group differences. Prior to describing and illustrating some of the methods for assessing item and test bias, we provide a few words about the terminology used for describing item and test bias.

Current Terminology

These days, item bias is typically referred to as differential item functioning (DIF) and test bias as differential test functioning (DTF). Early studies of item bias were stimulated by U.S. civil rights legislation in the 1960s. Test professionals wanted to identify test questions that minority groups (e.g., Blacks and Hispanics) responded to differently compared with the White majority group (Angoff, 1993; Cole, 1993). Angoff (1993) noted the following:

> These studies were designed to develop methods for studying cultural differences and for investigating the assertion that the principal, if not the sole, reason for the great disparity in test performance between Black and Hispanic students and White students on tests of cognitive ability is that the tests contain items that are outside the realms of the minority cultures. (p. 3)

Many assumed that biased items functioned against the minority group, that these items would be answered incorrectly more often by the minority (or focal) group than by the majority (or reference) group. Presumably, if these "biased" items could be identified and eliminated, test score differences between minority and majority groups would no longer occur.

In the late 1980s, the term *DIF* began to replace *item bias* in psychometric and professional testing circles. The reasons for this change probably had more to do with linguistics and politics than with psychometrics. The term *item bias* carried with it a negative connotation and was commonly associated with the notion of unfair, discriminatory testing practices rather than with its psychometric definition. Testing professionals felt it would be useful to separate technical, psychometric terms from those that may be politically and socially charged. Hambleton, Swaminathan, and Rogers (1991) write:

> Investigations of bias involve gathering *empirical* evidence concerning the relative performances on the test item of members of the minority group of interest and members of the group that represents the majority. Empirical evidence of differential performance is necessary, but not sufficient, to draw the conclusion that bias is present; this conclusion involves an inference that goes beyond the data. To distinguish the empirical evidence from the conclusion, the term *differential item functioning* (DIF) rather than bias is used commonly to describe the empirical evidence obtained in investigations of bias. (p. 109)

The examinations of DIF have been expanded beyond the early comparisons of groups that differ in terms of race and ethnicity. Nowadays, DIF analyses are frequently used to compare the performance on test items of groups that differ in terms of language, gender, disability status, and age. Researchers have also proposed that DIF analyses may help us understand the psychological processes involved in testing and "the subtle differences in content of a stimulus to which individuals react differently" (Cole, 1981, p. 1076).

Definition of DIF

An item *without* DIF may be defined as follows (Millsap & Everson, 1993):

$$\left\{\begin{array}{l}\textit{Probability of getting an item}\\ \textit{right, given a person's ability}\\ \textit{and group membership}\end{array}\right\} = \left\{\begin{array}{l}\textit{Probability of getting an item}\\ \textit{right, given a person's ability}\end{array}\right\}$$

(1)

In Equation 1, the lefthand side refers to the probability of answering an item correctly given a person's ability*and* his or her group membership, whereas the righthand side refers to the probability of answering the item correctly given a person's ability level *irrespective* of group membership. In essence, the equality in this equation means that the probability of answering an item correctly depends only on the person's ability. The fact that the test taker is a member of one group or another plays no role in the test taker's chances of answering the item correctly. If the equation holds true for an item at all ability (test score) levels, such an item is said to be functioning equally across groups. In other words, the item is said to have no DIF or bias and the item is considered invariant across groups that are examined. On the other hand, if the equality in Equation 1 does not hold, meaning that group membership increases or decreases the test taker's probability of answering correctly, then such an item is said to function differentially across groups and hence is designated as a biased item.

As mentioned previously, an analysis of bias at the item or test level usually involves two groups defined by demographic variables such as race or age (e.g., Blacks vs. Whites or old vs. young, etc.). Recent developments, however, have made it possible to examine bias or DIF across more than two groups simultaneously. In addition, the groups examined are not necessarily limited to subpopulations defined by physical characteristics. For example, the two groups considered could be employees and their immediate supervisors or peers, where ratings of employees by their supervisors and peers may be evaluated for DIF. In such an analysis, one would be interested in knowing whether supervisors and peers are giving the same performance ratings to employees with similar or identical work performance records. An analysis of this sort may help researchers identify rating bias by rater source. Another practical application of a DIF analysis is to establish the equivalence of translated tests. In this case, the language in which

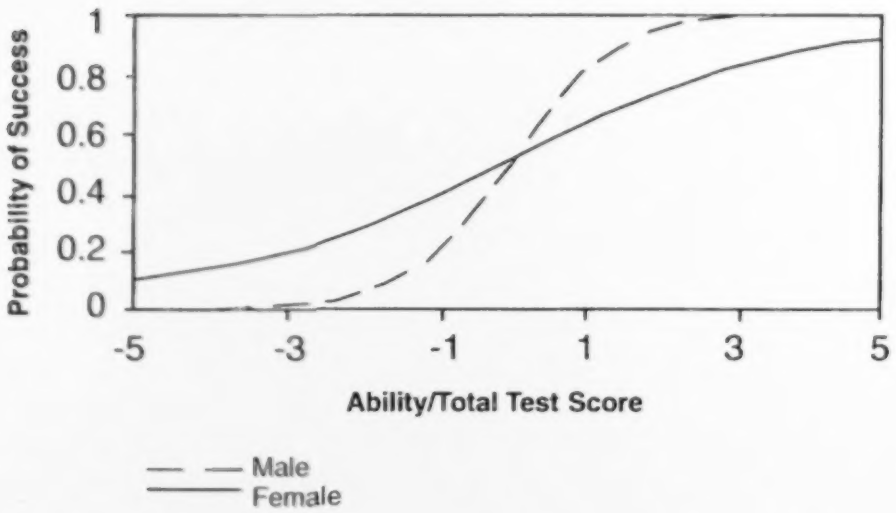
the test is administered (e.g., English vs. Spanish) would define the groups examined. This type of an analysis would provide practitioners with information about the quality of the translation beyond what a back translation would reveal (Ellis & Mead, 2000). Thus, for a DIF or bias analysis, the number of groups examined and the way groups are defined should depend on the test at hand and its intended use.

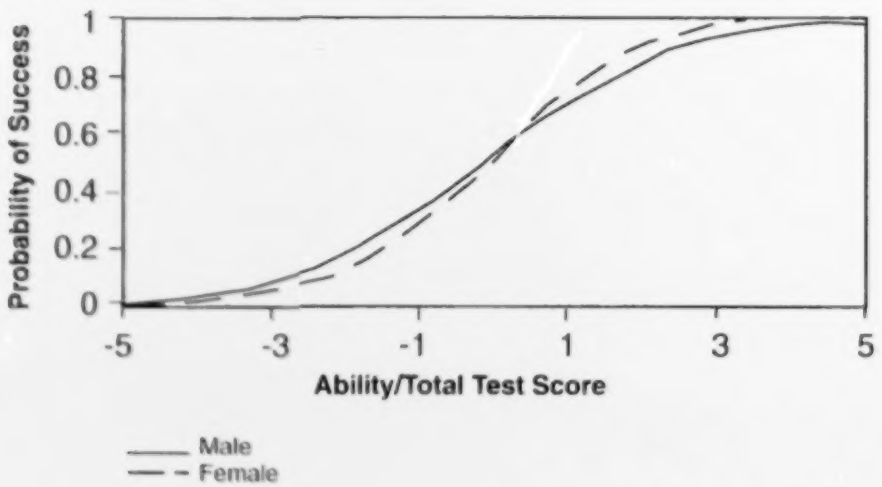
Techniques for Assessing DIF

There are many methods for assessing item bias or DIF. Some of these methods are based on classical test theory (e.g., the Mantel-Haenszel technique, logistic regression method, or SIBTEST), while others are based on item response theory (IRT; e.g., Lord's chi-square test, Raju's area measures, and the likelihood ratio test). Most of these methods provide similar information about DIF, but it is beyond the scope of this chapter to offer a description of these methods. Information about these methods may be found in Camilli and Shepard (1994), Holland and Wainer (1993), Millsap and Everson (1993), and Raju and Ellis (2002). We will, however, illustrate one of the IRT-based methods.

The method based on area measures is illustrated with two items, one with significant DIF, or bias, and the other with no DIF. Figure 1 shows separately for males and females the probability of getting an item right for a given ability or test score on a biased vocabulary item. The x-axis in this figure refers to the ability, or total test score, and the y-axis to the probability of answering the item correctly. When there is no bias, the probability graphs should be identical (or close to identical) for both males and females. The fact that these two graphs are different in Figure 1 implies that the item is biased, or has significant DIF. The graphs in this figure cross at an average ability score of 0.0 on a scale metric ranging from -5 to +5. Above and below this ability level, two persons with identical abilities will have different probabilities of success on the item. At the ability level of 1.00, the probability of success on this item is 0.82 for a member of the male group and 0.62 for a member of the female group. Even though two test takers have the same ability (i.e., 1.00), the individual from the female group has a lower probability of success than the individual from the male group; that is, the item under consideration favors the male group at this ability level. At the ability level of -1.00, the probability of success on the item also varies as a function of group membership, but this time the male group member has a lower probability of success (.18) than the female group member (.38), thus the item favors the female group. An

item of this type is said to have significant DIF, and the kind of DIF displayed in Figure 1 is called non-uniform DIF; that is, the type of DIF does not favor the same group across all levels of ability. In Figure 2, graphs for the focal and reference groups, although not identical, are very similar, indicating that the probability of getting an item right varies only as a function of an examinee's ability, not his or her group membership. These types of graphs are helpful in assessing not only the magnitude of DIF, but also where the significant DIF occurs. These graphs are also useful in exploring the reasons for significant DIF.





Challenges Ahead for DIF and DTF Analysis

During the last 20 years, we have made great strides in perfecting the methods, mathematical algorithms, and computer software required for assessing differential item and test functioning. However, many interesting and challenging questions remain unanswered. Some of these challenges are described in the following sections.

Understanding and Resolving DIF

Being able to identify DIF items represents a tremendous step forward for test developers, but the ability to identify DIF items raises new and challenging questions. Exactly why do some items have significant DIF? Furthermore, what should we do with DIF items once we have identified them as such? Test developers may choose to replace DIF items with new items, evaluate the new items for DIF, and repeat this process until all items in a test or scale are DIF (bias) free. But this an expensive and time-consuming process that may have negative consequences. For example, if a lot of DIF items are removed and replaced with new items, the construct assessed may be altered. Another approach would be to revise DIF items so that they no longer exhibit significant DIF and use these revised items in the final test or scale. The second method requires that the revised items be readministered to a new sample and reassessed for DIF. Both of these responses to DIF items implicitly or explicitly assume that the test developer can identify the source of DIF. Is this a valid assumption? Unfortunately, in most cases, the reasons for DIF or item bias are not evident. Thus, developing objective, testable methods for identifying the sources of DIF is one of the biggest challenges we face.

Editorial and Content Review of DIF Items

In developing tests, subject matter and editorial experts and members representing the groups under consideration (e.g., males and females, African Americans and Caucasians) usually review the questions. This panel may include sensitivity experts, but in most test development situations, a sensitivity review will have taken place prior to a DIF analysis. In a sensitivity review, items are examined for content that may be offensive or demeaning to members of a focal group. Most commercial test publishers have well-documented guidelines in place for use by their editorial staff members. These guidelines are designed to eliminate sexist and racist language and to avoid stereotypes about women and minorities. But, as Clauser and Mazor (1998) note,

"Sensitivity reviews are separate and distinct from DIF analyses—both are important, and neither can substitute for the other" (p. 32). Research indicates that it is very unlikely that experts will flag the same items that are statistically identified as having significant DIF (Engelhard, Hansche, & Gabrielson, 1990).

Following a statistical analysis for DIF, a committee of experts may be asked to develop hypotheses regarding the sources of DIF. Again, researchers have been disappointed to find that it is difficult or impossible to develop plausible explanations for the sources of DIF. At best, this exercise offers only a post hoc explanation of DIF that must be evaluated in future studies. Needless to say, more work is definitely needed to carefully articulate reasons for DIF in different content areas. Interested readers are referred to Camilli and Shepard (1994) and Holland and Wainer (1993) for further discussion on this topic.

Conclusion

Differential item functioning, or item bias, the assessment of how well two individuals with identical ability but different group membership perform on an item, is an important component of test and scale development. This definition of DIF does not imply nor does it require that the two *groups* under consideration be equal with respect to the construct being assessed (e.g., ability). The definition of non-DIF, or lack of bias, requires only that examinees with equal ability (or equal total test score) have the same probability of answering the item correctly irrespective of their group membership. There is a similar definition of differential functioning at the test level, called DTF. Assessing DTF is obviously important because decisions about examinees are usually based on their performance at the test level rather than at the item level. Although there are several known procedures for assessing DIF and DTF (i.e., item and test bias), many challenges still lie ahead for item bias research, especially in understanding the factors that contribute to item and test bias.

References

AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education]. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36,* 1067–1077.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25–33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ellis, B. B., & Mead, A. D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF Questionnaire. *Educational and Psychological Measurement, 60,* 787–807

Engelhard, G., Hansche, D., & Gabrielson, S. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347–360.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156–188). San Francisco, CA: Jossey-Bass.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").